Human Resources Mining for Examination of R&D Progress and Requirements

Sercan Ozcan [a, b, c], C. Okan Sakar [d], Metin Suloglu [e]

[a] Portsmouth Business School, University of Portsmouth, Portsmouth, UK

[b] Department of Engineering Management, Bahcesehir Universitesi, Istanbul, Turkey

[c] National Research University Higher School of Economics (Russian Federation), Moscow, Russia

[d] Department of Computer Engineering, Bahcesehir University, Istanbul, Turkey

[e] School of Computing, Leeds University, Leeds, UK

e-mails: {sercan.ozcan@port.ac.uk; okan.sakar@eng.bau.edu.tr; sc19ms@leeds.ac.uk}

Abstract

The amount of job advertisement data is rapidly growing, and this rich dataset is expected to have implications for the employment market, sector trajectories and the education sector. Most significantly, human resources (HR) data has never previously been examined with the lens of tech mining for science and technology analyses. Our study is the first to examine job advertisement data considering research and development (R&D) progress and requirements, and hereafter we refer to this as HR mining. The aim of this study is to use HR mining with the purpose of R&D and human capital intelligence using the job advertisement data of Turkey for the 2015–2017 period. The method of this study follows classification as part of the pre-processing step to determine R&D, engineering and high-tech industry-related job advertisements. Afterwards, we use clustering methods to identify areas where key human capital is required, and investments are made by R&D-oriented companies. The results show that it is possible to identify sector-oriented skill requirements and that the significance of the R&D skills varies. For the case of Turkey, we can clearly show the national human capital and R&D by identifying nine key clusters that indicate R&D progress and directions.

Keywords: Text Mining, Tech Mining, Human Resources Mining, HR Mining, Job Advertisement Analysis.

1 Introduction

Tech mining aims to extract useful information from technical resources with text mining methods for the purpose of investigating present and future technological developments [1]. Using these techniques to follow the newest technological advancements aids the improvement of business operations. The most common data sources tech mining is applied to include patents [2, 3], scientific publications [4, 5], and recently, social media data [6, 7] and website data [8]. The progress and developments in the tech mining field can happen by implementing one of the following: 1) new or adaptations of methods or algorithms for tech mining (i.e. applications of deep learning [9]), 2) use of existing methods in new types of data (i.e. HR data) and 3) persistent use of tech mining in similar data types to increase validity and reliability (i.e. application of tech mining to an emerging field).

With the aim of contributing to the tech mining literature and the community, and also considering the gap in the existing literature, we introduce the idea of 'Human Resources Mining', which will be referred to as 'HR Mining' in the rest of this paper. HR mining contributes to the field in three ways: 1) examination of a new type of data for this field, 2) building an end-to-end system to analyse HR data with the lens of tech mining, 3) practical findings and implications. To achieve these contributions, we use job advertisements with the aim of exploring the research and development (R&D)-based human capital investments, national human capital requirements in R&D and finally, future technological trajectories.

HR mining can be used in several areas based on academic, industrial and governmental needs. For academia, HR mining can be a way of developing an education strategy. For industrial actors, it can be a way assessment for the competitors, their own human capital or sectorial development. For governmental actors, HR mining can assist with informed policy development and assessment of national needs.

We believe that HR mining can be used for technology or R&D-oriented assessments and may have some advantages over patent and publication-based examinations, in addition to its many other potential applications. As HR investments in R&D-related personnel would be one of the initial steps before an organisation invested in certain R&D areas and technologies, HR mining can help identify technological and sectorial trajectories before they appear. Moreover, HR mining on R&D-related job advertisements can show a more dynamic picture and an earlier state of a country's R&D conditions compared to patent and publication analysis. In this study, we aim to provide extensive details on the HR mining methodology and show our steps and approaches. As part of the methodology, we apply text pre-processing, classification and clustering techniques on the HR job advertisement data taken from a substantial online recruitment website in Turkey. We only examine R&D, technology and innovation-related job advertisements.

Compared to existing studies, this study is original, due to the type of data that is being used and the methodological approach that is being implemented for the first time. This study brings a new way of examining sectorial, regional and national developments that were not possible without HR mining. HR data, which may include the job advertisements, vacancies and prevalent skills in the applicant pool, have been used for various purposes in the literature before. However, as detailed in Section 2.2, these studies mostly focus on matching job profiles with job advertisements or analysing the required skill sets in a specific domain and how those skill sets have changed over time. Our study is the first one that processes R&D HR data for the purpose of tech mining with an end-to-end analysis from the cleaning and pre-processing of HR data to the cluster map of keywords representing the clusters of R&D skillsets, their inter-relations and their relevant sectors. Thus, we propose to use a new type of data for tech mining.

From a technical point of view, the most important contribution of our paper is to propose an integrated pipeline of the most suitable methods to analyse HR data for tech mining. Some of the steps are either new or very rarely implemented in the tech mining field. First, we utilize classification algorithms as a pre-processing stage to eliminate the non-R&D job advertisements that will not provide useful information. In this stage, we combine multiple classifiers with the ensemble learning approach to obtain a more generalizable classifier. Second, considering that tech mining studies are mostly based on the use of short text data such as patent/publication titles or job advertisements as in our study, we use, to the best of our knowledge for the first time, the affinity propagation (AP) clustering algorithm to obtain the cluster map of HR data. The ability of this algorithm to deal with sparsity problem of co-occurrence matrices in short-text processing tasks has been shown before in the literature [10, 11]. We show that AR outperforms the most commonly used clustering methods in related studies.

The rest of this paper is organized as follows. Section 2 provides a background and a literature review on tech mining and relevant HR data examinations. Section 3 presents the description of the job advertisements dataset and methods applied throughout the analysis. Section 4 includes the experimental results and findings based on the interpretations of the clustering visual. Finally, we conclude in Section 5 with key findings, implications, limitations and suggestions for future research.

2 Background and Literature Review

In this section, firstly, tech mining literature is reviewed considering data sources and application areas. Next, HR and job advertisements-related data are analysed. Finally, the aim, objectives and the conceptual framework of the study are positioned based on the literature.

2.1 **Tech Mining for Tech Mining**

In this section, we will be reviewing the tech mining literature. The purpose of this section is to identify application areas, common methods and data sources of tech mining. By reviewing this literature, we are aiming to identify methodological and application-oriented gaps and weaknesses in the literature. Thus, we show that HR data it has never been used for the purpose of tech mining and some of the methodological steps can be enhanced using clustering and classification related approaches that are rarely or has never been used in this area. Centrality measures appear to be a more common approach compared to the clustering algorithms and classification approach it has never been implemented as a pre-processing step. The literature review in this section supports these points.

Accordingly, we followed the tech mining method to examine the tech mining literature in this field, as presented in Figure 1. Accordingly, we selected 215 articles and conference proceedings that used the term 'tech mining' or listed Porter and Cunningham's *Tech Mining* book [1] in the reference list. We acknowledge that we could have included the terms scientometrics, bibliometrics, text mining or patent mining as part of this review, but we aimed to study a highly specific group of tech mining studies as part of this section.

As shown in Figure 1, a distinct group of studies apply tech mining methods. Some studies aimed at examining national profiles, such as China, Japan and South Korea (many more can be seen in the visual) [12-14]. Some studies focus on emerging fields and technologies, such as nanotechnology [15], nanomaterials (i.e. graphene) [16], nanoparticles (i.e. dye-sensitized solar cells) [17] and nano-enhanced systems or materials (i.e. Nano-enhanced Drug Delivery (NEDD)) [18, 19]. Apart from the application areas, the data sources can be clearly seen as patent and scientific publication data. As a methodological approach, text mining [20] and bibliometric techniques [21] are mentioned with the tech mining studies. Recently, the term 'big data' has been appearing in tech mining studies due to its popularity [22]. The type of

analyses that it is mostly used for are foresight, technology and R&D analysis, technological roadmapping, decision-making process/models, competitive intelligence and collaboration (network) analysis [23–25].

Considering different methods in tech mining, there are two common approaches are used in the analysis step and for data visualisation. These are social network-based centrality measurements (i.e. betweenness centrality, closeness centrality, eigenvector centrality and degree centrality) or clustering analysis (i.e. k-means, DBSCAN and hierarchical clustering) [26]. Centrality measurements appear to be more common than clustering applications for tech mining or scientometrics. Clustering is an unsupervised method of finding groups in data where each sample in one group is more like others in the same group than to samples in any other group, under a certain aspect. Many clustering algorithms exist and differ from each other by the definition of a cluster and how to efficiently determine the clusters [21]. Specifically, the performance of clustering techniques generally depends on the initialisation, structure of the dataset, and choosing the correct dissimilarity measurement.

Cluster analysis has been used in many tech mining studies to explore the groups of patents, scientific publications, or keywords for scientometric analysis technological forecasting [21]. Some of the widely used clustering techniques in tech mining studies include hierarchical clustering [27, 28], k-means clustering [29, 30, 31], and density-based clustering [32]. These methods have also been used successively [32–34]. Especially as the hierarchical and DBSCAN algorithms do not require the number of clusters as a hyperparameter, they have been used before k-means clustering to determine an initial value for the number of clusters in the dataset. For example, Kyebambe et al. [32] used DBSCAN as a pre-processing step to identify the number of clusters and then used k-means to find the clusters. Litecky et al. [33] applied hierarchical clustering in the first step and obtained a number of skill set clusters and then used this number as an input to the k-means clustering algorithm.

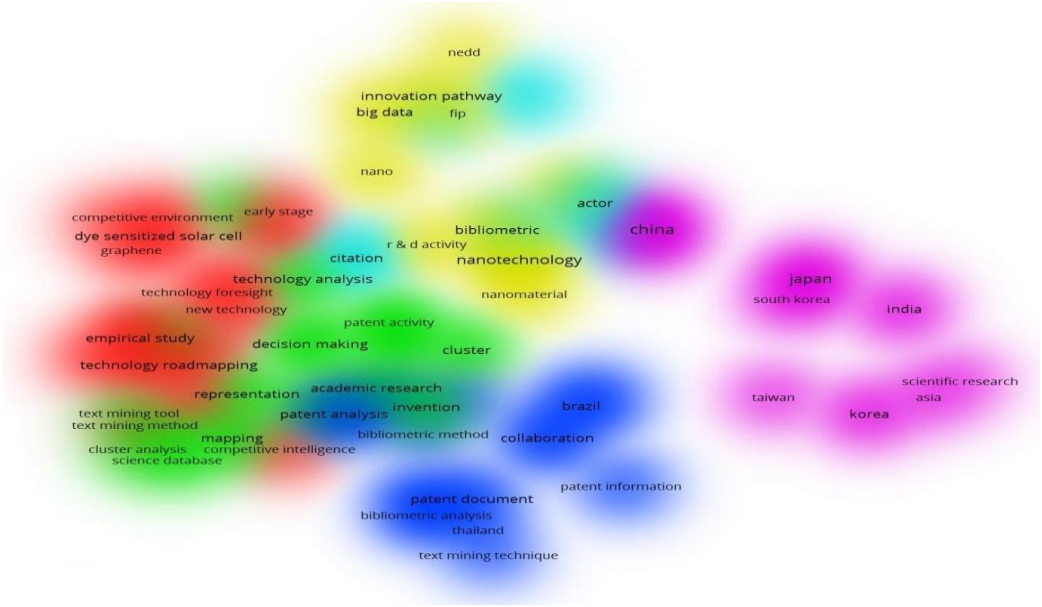


Figure 1: Tech mining for tech mining

In addition to these clustering algorithms, affinity propagation (AP) [35] is another clustering method that can be used on affinity (similarity) matrices. Other studies show that the AP algorithm significantly outperforms exemplar-based clustering approaches on various tasks both in terms of convergence time and the clustering quality [36]. AP has also been successfully applied to different text processing studies in the literature. For example, in [37], AP is used to construct concept maps from text documents, and the obtained concept maps are parallel to the outputs generated by domain experts. In [10], AP has been adapted to integrate the pre-known class information into the text clustering process. The authors showed that the modified AP outperforms k-means clustering on a benchmark text categorization dataset. Guan et al. [11] proposed a variation of AP for semi-supervised text clustering and showed that it outperforms k-means with better clustering results and faster convergence.

One of the main superiorities of AP compared to the other clustering algorithms is its ability to deal with the sparsity problem of co-occurrence matrices in short-text processing tasks. In [38], various clustering algorithms including k-means, singular-value decomposition and AP, were compared on short-text data collected from Twitter. The authors used two similarity metrics, cosine-based and Jaccard-based, with k-means and AP and compared their success using

clustering error as the evaluation metric. The best result was achieved using cosine-based AP. Kang et al. [39] addressed the drawbacks of the use of traditional topic modelling techniques in short-text clustering problems and proposed to apply AP to cluster tweets that contain URLs to news stories. Kang et al. [39] demonstrated the success of AP using the tweets' own URL as the actual cluster labels.

Apart from the use of clustering algorithms in the tech mining field, classification algorithms are not frequently used. Only a few studies address the adoption of classification algorithms to help automatically classify patent documents into their categories or to classify them based on their values [40-43].

Having reviewed the tech mining studies with an application point of view and considering the methodological approaches, we can clearly see that clustering and classification applications are weak in this domain. Hence, we identified some clustering and classification algorithms that are implemented scarcely in this domain and aim to implement different clustering and classification algorithms as part of our HR mining method.

2.2 **Use of HR Data with Machine Learning**

In this section, we examine the relevant literature that use HR data considering different methodological approaches and also the application areas, especially where HR data is being used for similar approaches that is aimed at our study. The first, we review the trends in job advertisements to see if big data scale approaches are required for this domain. Afterwards, we give an overview of the existing studies that apply machine learning techniques on HR data as we also propose to use HR data with machine learning and text mining techniques to identify technological and sectorial trajectories. Finally, we position our study focusing on where HR data has not been used previously.

According to the U.S. Bureau of Labor Statistics [44], there are more than seven million vacancies in the United States and nearly one million vacancies in the United Kingdom per year [45]. These job advertisements are a very rich source of data that can be used as a learning mechanism to understand the investments and the future directions of the sectors. The job listings can also be used as a feedback mechanism to inform the education sector and policymakers about the knowledge gap in the industry. The education sector can use job advertisement information as a way of developing or modifying education programs according to the needs in the sector. For instance, universities can develop undergraduate or postgraduate programs based on regional needs. Recruitment agencies can use HR mining as a way of identifying an opportunity, especially for the sector- or job-specific agencies. Governmental bodies can use HR mining to identify national needs and develop policies, programs and strategies accordingly. HR mining can be a great way for a business to examine its competitors' development, identification of regional capabilities and sectorial planning progress. On an individual level, HR mining can be used to identify relevant skillsets for a desired job role.

HR data, which may include the job advertisements, vacancies and prevalent skills in the applicant pool, have been used for various purposes, even if not for tech mining, in the literature. Most of these studies focus on matching job profiles with job advertisements. One of these studies [46] proposes a system that suggests matching job advertisements to software developers according to their activities in a social coding platform. Specifically, the goal of the study was to determine the similarity level between a job advertisement and a developer's profile in the related platform by combining many natural language processing techniques. Bal et al. [47] also proposed a matching system for eRecruitment that extracts terms from job advertisements, creates a lexicon of terms, and then uses cosine similarity to match job advertisements and resumes of the candidates. In another related study, Paparrizos et al. [48] used machine learning techniques to address the job recommendation problem. The authors

proposed a hybrid supervised learning model based on the combination of decision tree and Naive Bayes that aims to predict an employee's next job transition using their past job history as well as some key information associated with employees and institutions.

Some research efforts use HR data to analyse the required skill sets in a specific domain and how those skill sets have changed in time. One of the earlier works performed on the subject is that of Todd et al. [49], where the change of skill requirements for information systems (IS) jobs between the years 1970 and 1990 was reviewed using job advertisements taken from newspapers. Kennan et al. [50] have a similar study where IS positions in Australia were analysed using content analysis and clustering techniques. The authors claimed that the predominant cluster identifies the core skills new graduates are expected to have and found that technical knowledge requirements are present in a large proportion of advertisements. In another study, Litecky et al. [33] investigated computing job advertisements using cluster analysis methods to generate 20 job types and obtain the corresponding desired skill sets. For this purpose, the authors collected more than 200,000 job advertisements requiring a degree related to computing programs and applied the k-means clustering algorithm to map the skill sets to specific job definitions.

Focusing on the closest studies to the HR mining approach, we see that most studies examined the trends and the evolution of job advertisements using content analysis in this field [49, 51-53]. Using content analysis, the researchers manually coded advertisements, using different coders to increase the reliability of their studies. In a related study, Harper [54] reviewed 70 studies where librarian job advertisements were analysed as the main source of data and found that only three studies out of the 70 used automatic coding and that the use of automatic text mining approaches in job adverts was one of the key recommendations for future research.

Considering more advanced text mining techniques, Sanchez-Cuadrado et al. [55] and Marion et al. [56] used co-occurrence of words to examine the requirements for professional skills and

capabilities. Sanchez-Cuadrado et al. [55] illustrated concurrence of knowledge and competencies in job offers using a total of 1,020 job listings from a Spanish employment agency website for the period between the years 2006 and 2008. Their results indicated that some transferable competencies were required across many different job advertisements, such as a foreign language and the knowledge of information technologies. Their study was very useful in that it revealed the commonalities and differences across different job advertisements.

An examination of the methodological approaches of the studies in this field indicates that there are two weaknesses that need to be addressed: 1) job postings or other HR-relevant data are not being examined with the lenses of technology, R&D or foresight analysis, although a few studies use manual content analysis, and 2) as we approach the 'big data' era and the size of data increases, it is not a viable option to rely on manual involvements, and these large data sets need to be examined with fully automated or semi-automated approaches.

The existing research show that the size of job advertisement data increases and the analysis of this rich data source with advanced machine learning techniques offers many opportunities for various companies, sectors, and foundations. This section has demonstrated that there are limited number of studies where large scale HR data is being used and HR data has never been used for the purpose of tech mining. Hence, this section, 2.2 together with section 2.1 helps us to position our study considering methodological and contextual gaps.

2.3 **Theories and Models for HR Mining**

Considering the main purposes of HR mining, several theories and models support the foundations of the analysis and how the outcomes should be interpreted. These are 1)) Porter's Diamond model (PDM) [57, 58], 2) Core competencies (CC) [59-61], and 3) Dynamic capabilities (DC) [62-63].

In an organisational level, CC and DC can be seen as the key foundations to support the use of HR mining [64]. Human resource practices alone in dynamic environments, considering the DC, CC and the valuable, rare, inimitable and nonsubstitutable resources and capabilities (VRIO), cannot be adequate for a sustainable competitive advantage but that a human capital pool is a better strategy for a sustainable competitive advantage [64]. Hence, HR mining is crucial to retain the required human capital pool by benchmarking and assessing the changes in the market. Mobility of human resources makes it difficult for organisations to retain critical information, key talents and knowledge base within their companies, affecting a company's core competencies. For these reasons, HR mining can be used to assess changes in a sector for a company to stay competitive and dynamic.

At the national level, PDM [65] model support HR mining. PDM explains why one nation is more competitive than another in a particular industry [65]. PDM consists of four key determinants: firm strategy, structure and rivalry; factor conditions; demand conditions; related and supporting industries [65]. Factor conditions indicate the availability of resources and capabilities within a nation. Human capital is one of the key factor conditions why some organisations can perform better in a nation than another. In this determinant, advanced factors include skilled labour and specialist knowledge for organisational and national performance. Accordingly, an organisation may want to open a new branch or a factory in a location where the required human capital exists. To do that, HR mining can be the key approach to assess the opportunity and provide strategic direction. The government's role as a catalyst and challenger in PDM is important. A governmental organisation needs to make a national assessment to create a new policy and a new regulation, and at this point, HR mining could provide comprehensive results for the demand and the changes on a national scale.

Based on the abovementioned theories and models, HR mining's aim is described, and the conceptual framework is created in the following section.

2.4 **Research Aim, Objectives and the Conceptual Framework**

Tech mining examinations mostly occur based on patents and scientific articles using text mining, scientometrics, social network analysis and bibliometrics [1–5]. The findings of studies that use these data sources provide useful information only for the current key technologies because the related estimation techniques require historical citation data of the related patent documents [66, 67]. However, the data sources originating from the operations of companies, especially patent documents, generally only become available after the companies make investment decisions in the related areas, and the HR departments of these companies recruit the necessary workforce. Therefore, job advertisements can be considered as the initial indicators of potential technological developments. The information gained about upcoming developments in certain work areas earlier can be more useful than the information gained from comparably later outputs such as patents and scientific publications for technological forecasting.

Having reviewed the literature in the tech mining field with a methodological perspective, we identified that classification methods are not well implemented in this area. Only a few studies were found where classification methods are implemented, and there is a gap in studies that utilise classification methods as a pre-processing step to discriminate between related and non-related documents. In this study, we apply classification methods to determine and eliminate the non-R&D job postings to analyse only the R&D job postings. As part of the analysis, the tech mining studies mostly apply centrality measurements, social network analysis and topic modelling techniques along with text mining operations to analyse similar datasets. However, we found a very limited number of clustering applications in this field. Moreover, there was no study that applies HR mining for R&D job advertisements with the purpose of technology analysis.

Considering the methodological and practical gap, the aim of this study is to establish the first HR mining process with the purpose of examining a national scale technological developments and trajectories based on the R&D job advertisements data. Figure 2 displays the conceptual framework that illustrates the HR mining process and its application. The conceptual framework of HR mining is being supported by the relevant theories and models, as discussed in section 2.3. Accordingly, CC and DC models are used as the bases of capability assessment considering the required human resources with an aim of identification of key capabilities at sectorial level. PDM model is being used at national level, especially for the demand and factor conditions aiming to identify strategic directions and technological trajectories.

The objectives of this study are as follows:

- To establish the HR mining process by testing and identifying the most suitable approaches,
- To examine the clusters of R&D skillsets, their inter-relations and their relevant sectors,
- To analyse R&D job advertisements to understand national and sectorial requirements and the capability,
- To interrelate R&D job advertisements with technological and sector trajectories.

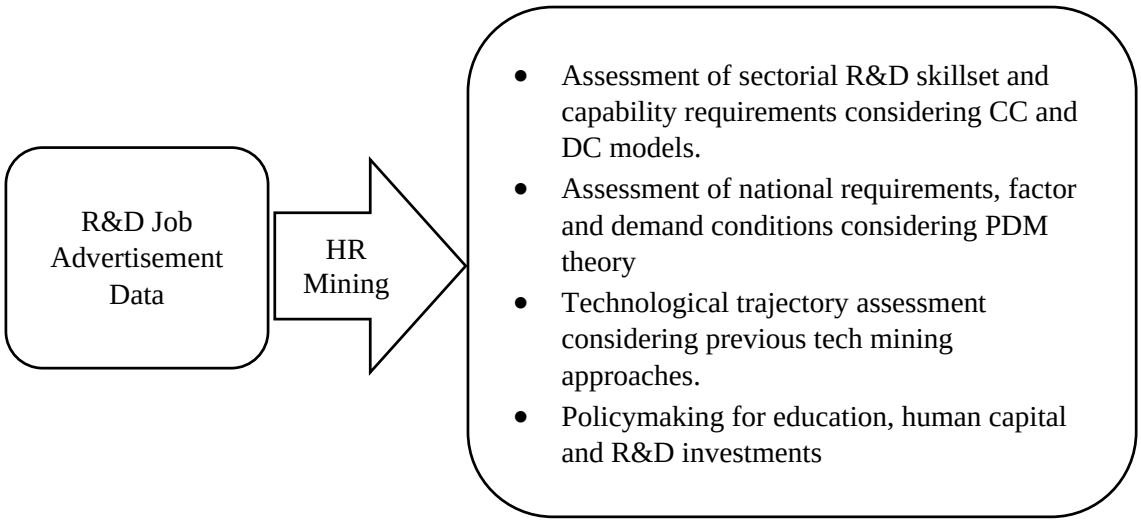


Figure 2: The conceptual framework for HR mining

3 Methodology

In this section, we describe the online recruitment dataset and give the details of the methods used to follow the proposed HR mining framework.

3.1 Dataset Description

The dataset used in this study consists of 3959 job advertisements gathered from Kariyer.net, the most substantial online recruitment website in Turkey. In addition to the descriptions of the jobs, the related company and sector information are also available in the dataset. Our own preliminary investigations of these job advertisements showed that some of the job advertisements were published in the wrong categories, and that a classification method based on the context of terms is a necessary key approach to identify the R&D-oriented jobs. As a result of the classification phase, the dataset size was reduced to 1035 R&D-specific job advertisements. We use only the job descriptions in the classification and clustering phases of the proposed system. After the visualization of the clustered keywords, company profiles and sector information are used together during the cluster-assisted in-depth interpretation phase of the system.

3.2 Methods

The steps of the proposed approach are given in Figure 3. Our study follows a mixed method of quantitative and qualitative approaches. Our quantitative approach is based on HR Mining, whereas our qualitative approach is based on the cluster-assisted in-depth interpretation by using job descriptions and company profiles available in the dataset.

In this section, we give the details of the methods used to follow the proposed approach. We first give the text pre-processing operations applied on the dataset, the text representation technique used to represent job advertisements, an oversampling technique used to deal with class imbalance problem, and the classification algorithms applied to eliminate the non-R&D

job postings from the dataset. Then, the clustering methods, including spherical k-means, DBSCAN, agglomerative hierarchical and AP, are briefly reviewed. We also present the details of the clustering evaluation metrics and a visualisation technique used to analyse the distribution of the important terms into the clusters.

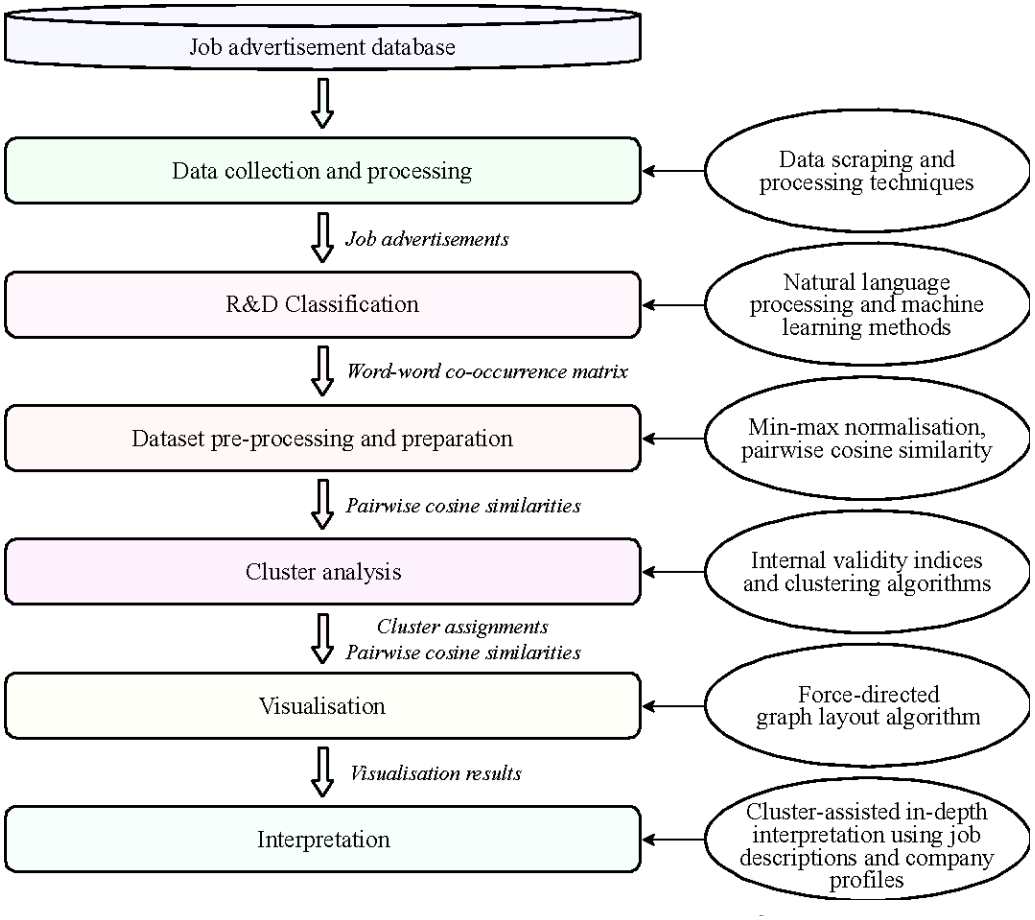


Figure 3: Steps of the proposed HR mining approach

3.2.1 Text Pre-processing and Representation

Text pre-processing and representation are important steps that are applied to text data before giving the data to the classifiers to ensure the machine learning models perform at their best. The steps of pre-processing are shown in Figure 4. To pre-process our data, we first convert all characters in our corpus to lowercase and remove all punctuation and stop words. Then, we tokenize all advertisements and obtain a list of words, some of which will be used to generate

our bag-of-words (BoW) representation. Afterwards, Porter's Snowball stemmer [68] is applied to all the words.

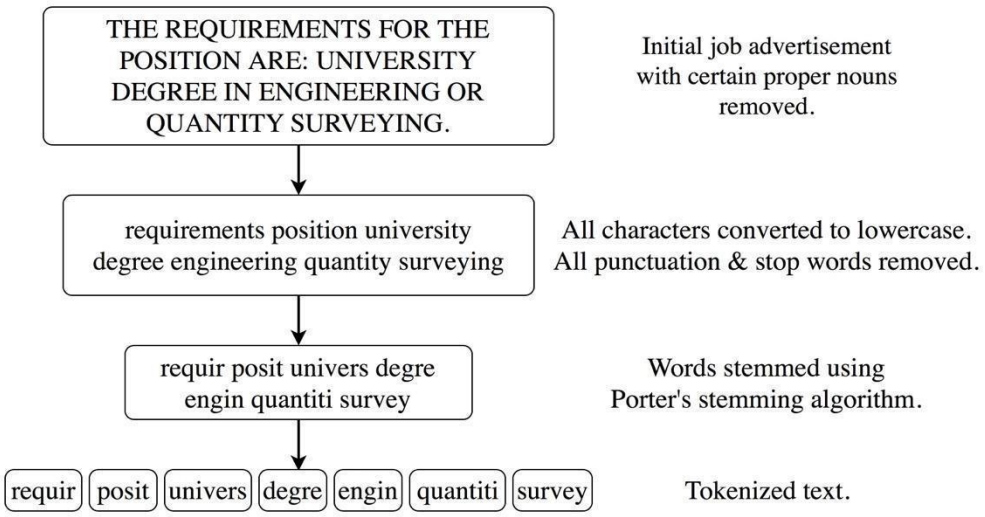


Figure 4. Pre-processing steps applied in the classification part of the system

The initial lexicon is extracted from the job advertisements by applying a set of text pre-processing operations, including stop-word elimination, stemming, and n-gram model representation. The obtained lexicon of words is not feasible to be used for a bag-of-words representation due to the 'curse of dimensionality' [69] phenomenon and high sparsity that worsens the performance of the learning and clustering algorithms. Therefore, a numerical statistic called term variance (TV) [70] is used as a filter to score and rank the words according to their importance in the corpus, and the top-ranked 225 words are selected. Then, for cluster analysis, we constituted the co-occurrence matrix of the selected 225 words, in which each word is represented in terms of its co-occurrence frequency with the other selected words. Each row and column of a word-to-word co-occurrence matrix specifies a word and the intersection of a row and a column shows the number of times one word appears in the same job advertisement as the other. We use cosine similarity to measure how closely words are linked together for all clustering algorithms.

The columns of our co-occurrence matrix are considered as features, and the row vectors are treated as the samples. As a pre-processing step of cluster analysis, we first scale the features by applying min-max normalisation to each feature vector. Then, an affinity (similarity) matrix is obtained by calculating the pairwise cosine similarity between samples. Performing min-max normalisation on each feature is important before performing these pairwise calculations because as the scale of one of the features tends to infinity, the cosine similarity between each pair of vectors approaches 1. Furthermore, all elements of the affinity matrix are in a range of [0, 1], and therefore, the measurements can easily be converted between distance or similarity by subtracting 1 from each cell.

In our study, we use a BoW [71] representation of job advertisements using term frequency-inverse document frequency (TF-IDF) values. The TF-IDF statistic is one of the most common text representation techniques used to represent the importance of any term in a document. It is also used to represent scientific documents in tech mining field [72]. We use the scikit-learn API for Python [73] to apply the TF-IDF feature extraction technique on our dataset. After text pre-processing operations, we extracted unigram and bigram representations of all words and fed a subset of these representations to the classification algorithms as detailed in Section 3.2.

3.2.2 Oversampling Using SMOTE

In our preliminary experiments, we identified that there is an imbalanced data problem where the number of non-R&D job postings is higher than the R&D job postings. Classifiers may struggle to perform well when there is imbalanced data. To fix this issue, we use the Synthetic Minority Oversampling Technique (SMOTE) [74] to generate new samples for the R&D class, which is the minority class in our dataset, before giving the job postings to the classification algorithms. In SMOTE, the new samples are created using a few minority class samples that are in the neighbourhood of the given sample. The most common value for the size of the

neighbourhood, *k*, is 5 [75], which is also used in our experiments. We used the Euclidean distance metric to determine the nearest neighbours of a given sample. The steps of the SMOTE algorithm used in this paper can be given as:

Input: Dataset containing minority class samples **x**; SMOTE parameter *N*%; Number of nearest neighbours *k*
Output: Synthetic samples S
$S \leftarrow \emptyset$
$T \leftarrow |x|$ *// Number of samples in x*
$N = \lfloor N/100 \rfloor$

for $t \leftarrow 1$ **to** T **do**
  Find *k* nearest neighbours of sample x_t
  for $n \leftarrow 1$ ***to*** N **do**
    Randomly select one of the *k* neighbours, call it $\underline{x_t}$
    Select λ uniformly from the range [0,1]
    $x_{new} = x_t + \lambda\left(\underline{x_t} - x_t\right)$
    Append x_{new} to S
  end
end

3.2.3 Classification

We use classification algorithms to eliminate non-R&D job advertisements from the dataset. Three different algorithms are trained for the classification task: logistic regression, support vector machines with two different kernels, and extreme gradient boosting algorithms. We also combine the predictions of the classifiers using an ensemble learning approach to obtain a more generalizable model.

Logistic regression is a simple but effective statistical model used for the classification of linearly separable data. We have included logistic regression in our experiments because different variations of logistic regression have successfully been applied in many text classification studies [76, 77]. Support vector machines (SVM), another commonly used classifier in the related domain, aims to find a hyperplane that not only separates the classes but also has a maximal distance to the samples that are hardest to classify [78]. We used SVM

with linear and radial basis function kernels in our study to discriminate R&D and non-R&D job postings.

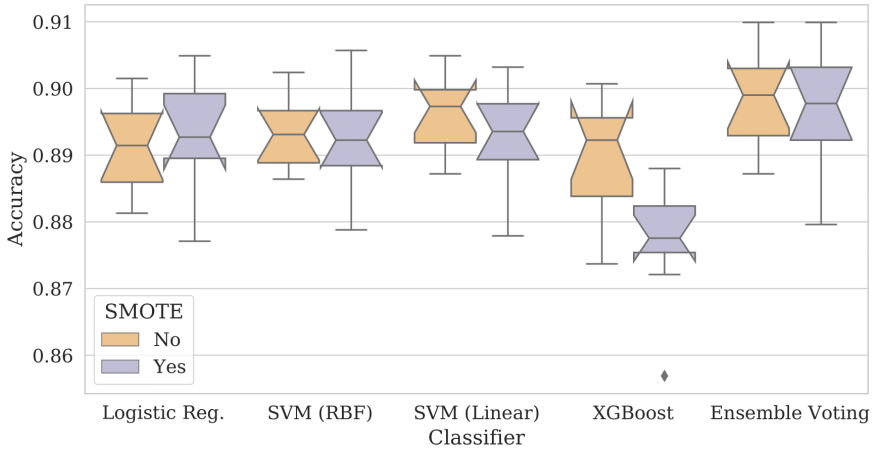
Boosting algorithms combine several models in an additive manner to obtain a classifier with minimal loss [79]. In boosting, many classifiers are generated sequentially, and the weights of the training instances fed to classifiers as input are changed so that the latter models give more importance to the samples that are incorrectly classified by the previously trained, weak models. The goal is to strengthen the predictive power of the final model. The models that use the gradient of the loss function to identify the drawbacks in each iteration are called gradient boosting machines (GBMs) [80]. Extreme Gradient Boosting (XGBoost) [81] is a tree-boosting library developed for scalable and efficient gradient boosting. It has become a popular framework due to its success and computation speed in many machine learning challenges [81]. Therefore, we use XGBoost to predict the type of job advertisements as a part of our final ensemble model.

Ensemble learning is based on combining the predictions of multiple individual learners with the aim of improving the generalisation ability of the final model [82]. The success of ensemble learning has been shown in many supervised and unsupervised learning problems [83]. In our experiments, we used the majority voting strategy to combine the predictions of the individual classifiers. We allow four classifiers to vote for the class. In the case of equality after voting, one of the classes is randomly picked as the final prediction.

3.2.4 Classification Results and the Selected Classifier

Figure 5 shows the box-and-whisker plots of the accuracies (top figure) and the F_1 scores (bottom figure), respectively, obtained with each of the classifiers to discriminate between R&D and non-R&D job postings. As seen in Figure 5, the highest (median) accuracy is obtained by combining the predictions of all classifiers using ensemble voting. However, we

also see that the notches of all classifiers, including the versions with and without SMOTE, overlap with each other, indicating that, at the 5% significance level, the difference between the accuracy of these classifiers is not significant. The only exception is for the versions of XGBoost with and without SMOTE that show the overall accuracy of XGBoost decreases when SMOTE is applied on the training set to obtain a class-balanced dataset. The highest median of the F_1 score is achieved with ensemble voting. We see that applying SMOTE does not affect the F_1 score for XGBoost, which indicates its robustness to imbalanced data. In fact, it is clear from Figure 5 that the synthetic minority class examples have decreased the accuracy for XGBoost. Figure 5 has the highest variance, which shows that the model is not robust to different subsamples of the same set of samples. For all other methods, the F_1 score values have significantly improved after SMOTE, indicated by the higher accuracy scores. The maximum F_1 score was obtained with the ensemble voting technique and SMOTE, while the minimum values of the F_1 scores are also larger with the voting technique, both with and without the application of SMOTE. Accordingly, the ensemble voting technique with SMOTE has been found to be the most successful algorithm in labelling the job postings as R&D and non-R&D, eliminating the non-R&D job postings before passing the data to clustering algorithms for HR mining analysis.



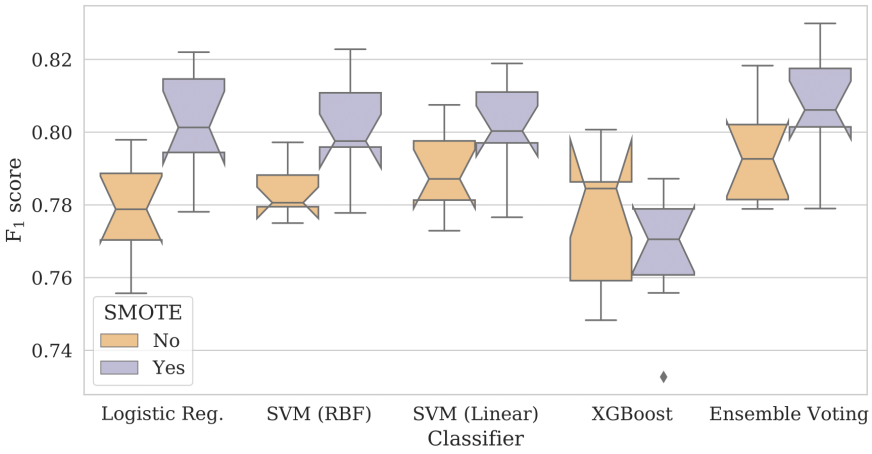


Figure 5: Accuracy (top figure) and F1 score (bottom figure) of classification algorithms with and without SMOTE.

3.2.5 Clustering Methods

In our experiments, we use four clustering techniques to identify groups of keywords in the job advertisement data. The details of these clustering algorithms are specified below.

The k-means clustering algorithm has successfully been used for cluster analysis in the tech mining field [29-31]. Classical k-means aims to minimise the sum of squared Euclidean distances between samples and the mean of their respective cluster:

$$E = \sum_{i=1}^{C} \sum_{x \in C_i} \left|\left|x - \mu_i\right|\right|^2 \qquad (2)$$

where C is the number of clusters and μ_i is the mean of cluster i. A global minimum of this function may not be reached depending on the starting position of the cluster means. K-means is generally used with Euclidean distance and may fail to model data that are not isotropic and have clusters with differing densities [84, 85]. Spherical k-means [86] is a slight variation of the traditional k-means algorithm. It uses cosine similarity to assign a cluster label to each of the data points. The initial cluster centroids may be initialised by selecting a random subset of data points, or by employing a more careful approach when selecting the initial points, such as by using k-means++ [87].

The k-means partitioning clustering algorithm divides samples into non-overlapping subsets such that each sample is exactly in one cluster. In hierarchical clustering, a set of nested clusters are created by sequentially merging or splitting existing clusters [88]. We used agglomerative hierarchical clustering because it is more common than the divisive approach in text clustering [89]. In agglomerative clustering, the process starts by assigning each data point to its own group and then the data points are successively merged in a recursive manner. The clustering process can be stopped when desired number of clusters is obtained. A similarity or distance metric is used to choose the next cluster(s) to be processed.

Density-based spatial clustering of applications with noise (DBSCAN) [90] is a clustering algorithm that is specifically designed for noisy data where the expected clusters are of different shapes and sizes. Instead of assigning each sample to a cluster, DBSCAN marks samples that are not in the local region of any other sample as noise as a pre-processing step and eliminates such samples before running the clustering process. Only the remaining points, which are marked as core points or border points during clustering, are assigned into a group. One of the important advantages of DBSCAN is that it does not require the number of clusters to be determined before the clustering process. Instead, two different parameters, Eps, which specifies the maximum distance between two samples for them to be considered neighbouring points, and minPts, which indicates the minimum number of samples around a point for it to become a core point, should be specified. DBSCAN is generally extremely sensitive to these parameters, often fails to identify clusters with differing densities, and does not work well on high dimensional datasets [91, 92]. On the other hand, the algorithm works well when the data contain clusters with arbitrary shapes [92].

We also used the AP [35] technique due to its success in previous text clustering tasks [36-39]. AP passes messages between points to find clusters in the data. It aims to identify one sample in each cluster as an exemplar point that best represents all other points in the same cluster. The

algorithm simultaneously evaluates each of the data point in the dataset as candidate exemplars until the stopping conditions are met and exemplars and clusters are determined. Two measures are used for each pair of points that are exchanged during the execution of the algorithm. $r(i,k)$ is called the 'responsibility' and indicates the suitability of point k serving as an exemplar for point i compared to all other candidate exemplar points for i. The 'availability' $a(i,k)$ is sent from k to i and represents to what extent point k is suitable to be chosen by point i as its exemplar. Each point k also starts with a 'preference' value, $s(k,k)$, that suggests how likely it is for that point to become an exemplar point. $s(i,k)$ indicates the similarity between point i and k. By applying certain update rules for these values and iterating until convergence, exemplar points and clusters can be obtained. More details of the method can be found in [35].

3.2.6 Clustering Evaluation

An important question that should be addressed in a clustering task is the optimal number of clusters required to represent the data [93]. The groupings of samples may not be immediately obvious, or there may not exist any grouping in the dataset at all. Especially in higher dimensions where viewing the distribution of the samples is impossible, different methods should be used to gain intuition about the dataset. A large number of clustering validity indices exist for the purpose of measuring the quality of the obtained clustering; however, because we have no information on the correct grouping of the words, we are limited to using internal clustering validity indices that do not use external cluster information in their calculation. Alongside comparing clustering quality between different algorithms, internal indices are also usually used to determine the number of clusters [93]. We use two widely used clustering validity indices to compare the algorithms: the silhouette width [94] and the Dunn index [95]. The silhouette coefficients are always in the range [−1, 1], and a larger value of the overall average score indicates a clustering where clusters are better separated. The Dunn index takes

a value greater than 0. As with the silhouette width, a higher Dunn index indicates a better-quality clustering result.

3.2.7 Visualisation

We visualise the clustering of co-occurrence matrix obtained from the R&D job postings by treating our data as a complete graph. The nodes in this graph are the data points, and edge weights represent the cosine similarity between points. In order to give our graph better structure and reduce edge crossings, we simplify the graph by removing edges with a weight value under a certain threshold. The process of removing edges and obtaining a simplified subgraph is often referred to as 'edge filtering' and has been shown to be effective in improving the layout of certain networks [96, 97]. In our case, edge filtering allows the nodes to move around more freely when laying out the graph and emphasises the strong links between similar nodes. After reducing edge clutter, we lay out the nodes while giving importance to edge weights by using the force-directed ForceAtlas2 [98] graph layout algorithm.

3.2.8 Interpretation of HR Mining Results

Our HR mining results using R&D job advertisements indicates a cluster of advertisements that needs in-depth examination based on the visualisation results. Each cluster and interlinkage of terms indicates a group of job advertisements that need to be examined. Accordingly, an in-depth examination of the job descriptors followed, focusing on the prospective employee's role and the required skill set, the project details, the sector information, the company details and the technologies that the new hire will be dealing with. A collective examination of each cluster led to the identification of sectoral needs and technological trajectories. Technological trajectories are identified mainly based on the collective details of many projects and the employees' expected duties.

4 Results

4.1 Clustering Evaluation Results

Figure 6 shows the change in the overall average silhouette width (the figure on the left) and Dunn index (the figure on the right) on our dataset with a differing number of clusters for four clustering techniques. The DBSCAN results shown are calculated using a constant minPts parameter of 4 and by not considering the samples classified as outliers. K-means++ was used to initialise the centres for spherical k-means. The complete-link merge criterion was used for agglomerative hierarchical clustering.

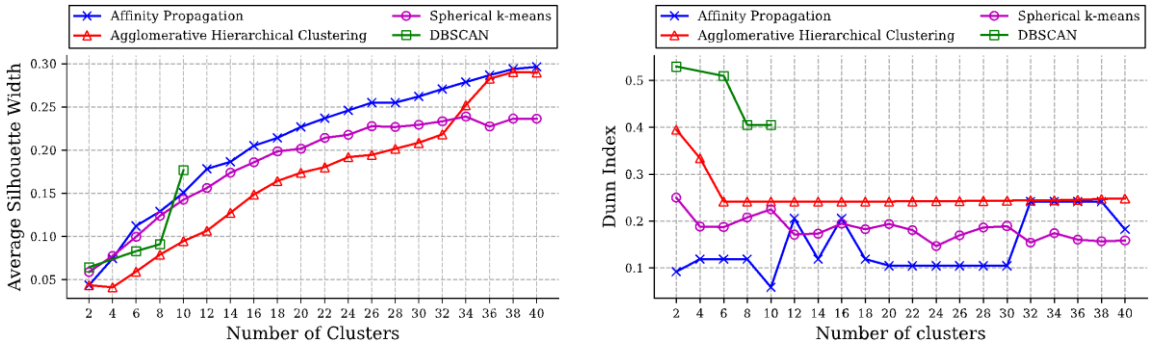


Figure 6: Values of **(left)** the Average Silhouette Score and **(right)** Dunn validity index versus the number of clusters

As seen in Figure 6, in general, affinity propagation (AP) is the top-performing algorithm according to the silhouette width measurement. Although DBSCAN gave a higher silhouette width for 10 clusters than AP, it marks a large subset of the keywords as outliers as a pre-processing step and does not assign such points to a cluster. Using 10 clusters to examine our dataset, DBSCAN classifies approximately 34% of the samples as outliers as seen in Table 1. Using six clusters, DBSCAN classifies approximately 22% of the samples as outliers, and about 62% of the samples are grouped into a single cluster. These trends also continue for different values of the minPts parameter. The small uneven clusters found by DBSCAN inadvertently give the method a higher score with the clustering validity indices, and the

clustering results obtained for the corresponding settings are not useful to evaluate the R&D fields present in the job advertisement dataset. The results in Figure 6 show that spherical k-means gave the second-best results in general after AP, according to the average silhouette width.

Table 1: The number of samples in each cluster for the AP and DBSCAN algorithms

Method	**Number of samples in each cluster**											**Total**
	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	Noise	
AP	50 (22.2%)	46 (20.4%)	25 (11.1%)	20 (8.9%)	16 (7.1%)	15 (6.7%)	15 (6.7%)	15 (6.7%)	12 (5.3%)	11 (4.9%)	-	225 (100%)
DBSCAN	80 (35.6%)	18 (8.0%)	15 (6.7%)	7 (3.1%)	5 (2.2%)	5 (2.2%)	5 (2.2%)	5 (2.2%)	4 (1.8%)	4 (1.8%)	77 (34.2%)	225 (100%)

When the clustering algorithms are evaluated in terms of the Dunn Index (see Figure 6), the clusters given by DBSCAN have higher Dunn Index values than the other algorithms. However, as we have noted above, DBSCAN discards a sizeable subset of the job advertisements in its initial pre-processing step. Therefore, we exclude the DBSCAN clustering results from our analysis. While AP gave a higher silhouette width than k-means and hierarchical algorithms, the highest Dunn Index value is obtained with hierarchical clustering. While the validity scores used here give a rough idea about the quality of a clustering, the results may be misleading in a sense that they do not correlate well with the human perception of a ‘good’ clustering result. Lewis et al. [99] statistically investigated the relation between how humans view good groupings and the results obtained by using clustering quality measures and found that the larger silhouette width scores have a higher correlation with the clustering results chosen by humans as the best compared to other clustering measures, such as the Dunn Index. This result also corroborates with our evaluation that AP groups more meaningful words together while also performing the best according to the silhouette width. Therefore, we give more importance to the silhouette width measurement when assessing the clustering qualities and present the analysis of the job advertisement clusters based on AP clustering results.

In clustering, another important point is to determine the number of clusters present in the data. In Figure 6, it is seen that for AP there exists an 'elbow' point in the silhouette width plot (the figure on the left) when the number of clusters is 12, and the plot of the Dunn index (the figure on the right) has a local maximum at 12 clusters. Another slight elbow point exists at six clusters for AP that can enable us to analyse the distribution of the job advertisements from a different perspective by providing a higher-level grouping. We provide the visual distribution of the samples for 12 clusters and present further analysis of the R&D activities explored by the corresponding clustering.

4.2 **HR Mining Results**

We follow the results of the average silhouette width plot shown in Section 4.1 and visualise the results of the top-performing clustering method, AP, for 12 clusters in Figure 7. The figure shows nine clusters labelled with letters A-I (three clusters are not considered separately as two clusters were part of another major cluster, and one cluster was too small to be considered).

As seen in Figure 7, the keywords are highly related to the R&D, technology and innovation-related job advertisements, which shows that the classification techniques detailed in Section 3.2, applied as a pre-processing step to eliminate the non-R&D advertisements, fitted well to the problem. The next step of the proposed HR mining framework shown in Figure 3 was clustering of the co-occurrence matrix obtained from the R&D advertisements. As seen in Figure 7, the clusters obtained with the top-performing clustering method revealed groups of related keywords. Note also that the similar clusters are placed close in the cluster map. The detailed interpretation of the clusters is given below. For visualization, we filter out edges with a weight less than 0.15 and colour the nodes based on the clusters found by the algorithms. The edge weight threshold was chosen such that every node is connected by an edge to at least one

other node, and the resulting graph is connected. Both the scaling factor (k) and edge weight influence (δ) parameters of the ForceAtlas2 algorithm are set to 2 for our experiments.

Cluster A shows the required skillsets, R&D investments and trajectories mostly related to software development. This cluster mostly includes skillsets related to the end-to-end development and management of software, from design to implementation, with new features. Cluster B, which includes more specific skillsets such as Javascript, Android, and iOS, mostly focuses on the development of web-based and app-based applications. Cluster C is a central cluster that includes keywords related to various steps of product development life cycle such as requirements analysis, research, algorithm design, prototyping, development, documentation, and testing. Therefore, cluster C can be considered as the heart of R&D activities. Cluster D is an interim skillset where software- and hardware-related areas merge and are mostly related to the embedded engineering, design, and prototyping. Cluster D shows significant overlap with Cluster C due to common skillsets in the design and prototyping process. Specifically, it illustrates required skillsets related to the automotive industry such as powertrain engineering and transmission.

Cluster E is related to the telecommunication sector and the required skillsets of network engineers. In this cluster, LTE, 3G and cell are some of the words that are linked to telecommunication skills, and network security, server and wireless are some of the words that appear for networking skills. Cluster F shows the recently developing big data and data science related skills. Cluster F is overlapping with Cluster E due to the common skillsets and requirements such as the use of big data infrastructure in telecommunication, network and databases. As seen in the map, Python keyword is in the intersection of clusters B, E, and F since it is a very increasingly popular scripting language used in various fields such as big data analysis and processing, open source software development, networking, and network security. Cluster G is about manufacturing systems such as automation, robotics and production lines.

As the manufacturing process is highly related to the operations management, terms such as lean, kaizen and TPM appear in this cluster. Cluster H is related to energy management and renewable energies. Accordingly, the terms such as solar and wind turbine appear in this cluster. Cluster H also has terms such as grid, supervisory control and data acquisition (SCADA) and power plant that indicate energy management systems. Some of these requirements are found to be related to smart grids as the future of energy management. Cluster I is about quality control and assessments. Accordingly, the terms such as quality standards, international standards and environmental safety appear in this cluster. Cluster I is close to Cluster G and H because both fields require high involvement in quality assurance, auditing and environmental safety. Our database shows that there are many job advertisements where energy management-related roles require knowledge of environmental impact assessment and auditing processes.

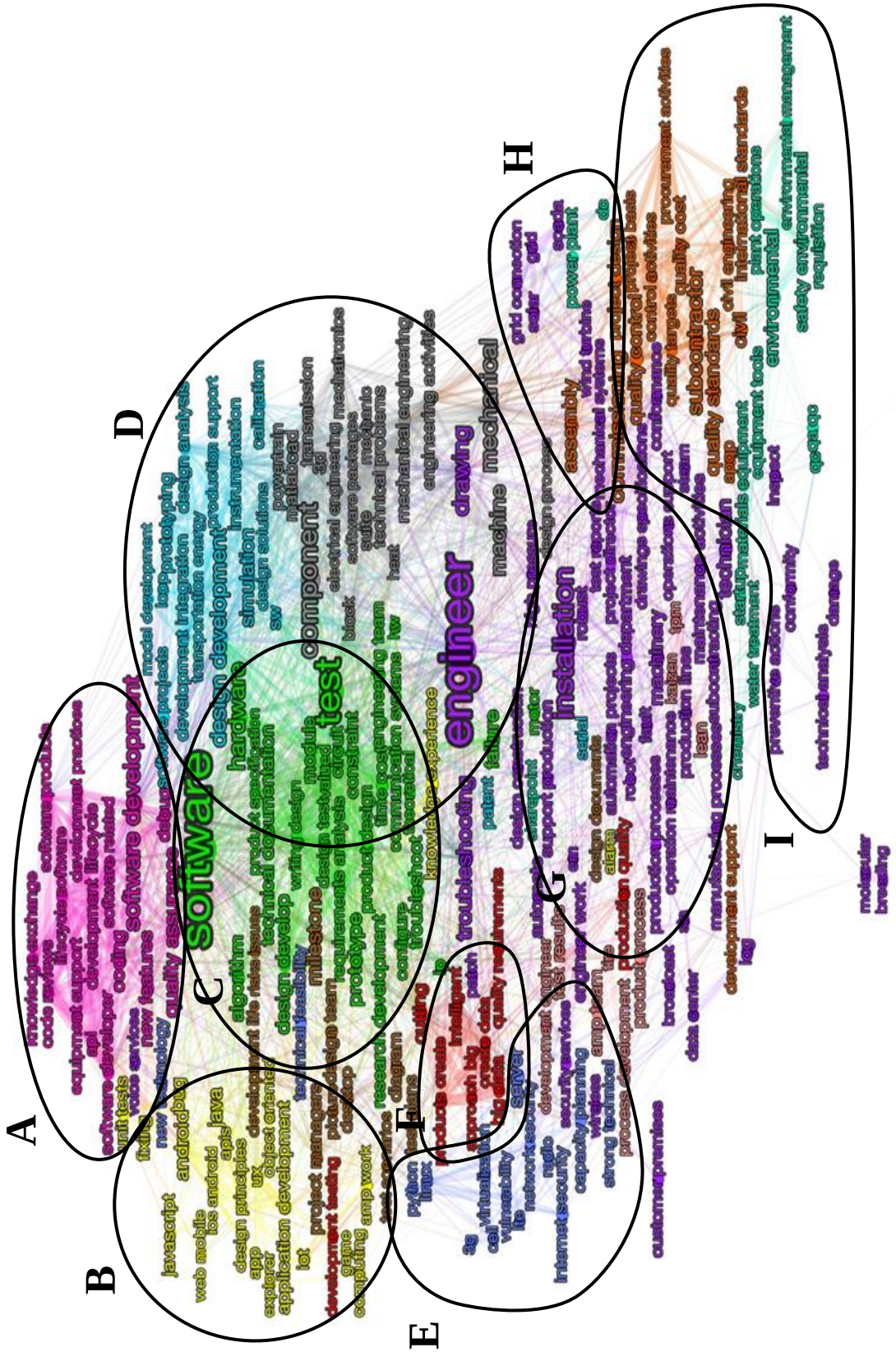


Figure 7: HR mining visualisation using clustering

Table 2 is created to relate the clusters of R&D skillsets and human capital requirements to the sectors and finally to map the national capability requirements with the sector and national R&D trajectories. Apart from Cluster H and I, all clusters have promising technology and innovation trajectories due to the investments in human capital and actual R&D. Table 2 is important because it shows: 1) key sectors that invest in R&D and that have continuity in their progress and investments, 2) sectoral and national R&D requirements and hence development (the data provider confirmed that the majority of these job advertisements had a successful recruitment), and 3) sectoral and national R&D trajectories as job descriptions directly or indirectly indicate R&D projects, and in some cases, the R&D trajectories are clear based on the company descriptions (i.e. the goal and what they are trying to achieve). Some job descriptions are limited, but our clustering approach indicated the group of skillsets and R&D and allowed us to see a bigger picture from a sectorial and a national point of view.

Table 2: Sector and national progress and requirements based on HR mining

Clusters	**Relevant Sectors**	**Required Roles & Skillsets**	**R&D Trajectories**
Cluster A – Software Development	Software & IT, Retailing, Telecommunication, Home Appliances, Banking & Finance, Tourism.	.NET, Java, J2EE, C, C++, C#, Objective-C, SQL Server, Oracle, SAP, GUI Development, HTML5, JavaScript, CSS3, AJAX, SOAP, REST, XML, JSON, CSS3, SASS, Photoshop, JSF, AngularJS, Apache etc.	Improvements or development of new software for sectors such as telecommunications, banking and tourism. Software-based new service developments and solutions. Software support for ERP, CRM, BI systems and e-business etc. Maintaining and using existing software.
Cluster B – Web and App Development	Software & IT, Banking & Finance, E-business, Electronics, Telecommunications, Marketing & Advertisement.	iOS, Android, Objective-c, Java, Xcode, C++, SOAP, XML, JSON, REST, WSDL, XSD, RAML, Linux, UNIX, Apache, HTML, Object	Mobile app and website development for a variety of sectors. New business development, especially for e-businesses.

		oriented programming, SQL, UI/UX, etc.	Marketing and advertisement-related support and development.
Cluster C – Product Development & Design	Home Appliances, Automotive, Defence & Military, Supplementary & Parts Manufacturing, Electronics, Telecommunications, Energy, R&D Institutions.	AutoCAD, CATIA, ProEngineer, SOLIDWORKS simulation tools, MATLAB, PCB design tools, Altium Designer, PADS, 2D/3D mechanical design, 3D printers & imaging, Finite element analysis, LV/MV, thermodynamics, HVAC systems, Hydraulics, Multi-body dynamics CAE, Hypermesh, Nastran, Patran, etc.	New product development and innovations in a variety of industries, especially home appliances, automotive parts, defence equipment and devices. Many international-project and global-company-oriented roles and development. Energy sector appears to be developing its own products and systems.
Cluster D – Embedded Systems	Defence & Military, Electronics, Support & Service, Telecommunications, Payment Systems and Security.	C, C++, Linux, Object-oriented design and programming, Source control management (SVN, GIT), Circuit board design, Microcontrollers, PCB design, MATLAB, Simulink, Stateflow, Raspberry Pi, TCP/IP, etc.	Many defence-oriented roles, especially in telecommunications. Payment systems-related expertise and development are highly apparent.
Cluster E – Telecommunication & Network	Telecommunications, Network, Security and Defence & Military	Object-oriented programming, Cloud architecture, DevOps, TCP/IP, SQL, NoSQL, International information security standards, Systems and network management, Firewall, Antivirus, IPS, Network security, Database management systems, Penetration test, Stress test, Encryption, Firewall, Intrusion prevention systems, Vulnerability scanning systems, DLP, Log management.	Research and development for secure network and telecommunication services. Global system for mobile communication (GSM)-related progress and developments. Secure telecommunication systems for the defence industry.
Cluster F – Big Data and Data Science	Software & IT, Electronics, Telecommunications, Banking, Finance, Insurance.	Python, Scala, R, GNU/Linux Hadoop, PIG, HIVE, Spark, H2O, Dato, Jenkins, Network intrusion detection, OpenStack, VMware,	Big data-based, real-time applications and solutions. High performance, real-time distributed

		KVM, Cloud computing SDN &;NFV, Open vSwitch, OpenDayLight, OpenFlow, ONOS, OpNFV, ONF certification, QoS, RabbitMQ/AQMP, Hazelcast, High performance, real-time distributed programming, Large scale system architecture, E2E solution architectures, Cassandra, RamCloud, HazelCast, HDFS, MapReduce, MapR.	programming for telecommunications. New database architectures and solutions on big data scale.
Cluster G – Manufacturing Systems and Operations Management	Electronics, Supplementary & Parts Manufacturing, Industrial Equipment, Logistics, Retailing, Textile, Home Appliances, Automotive, Defence & Military.	Industrial automation, Use of robotics and sensors, Simulators, Real-time motion systems, Vehicle modelling, Safety standards (i.e. ISO/TS standards), Process analysis, Process development, Project management, Lean management, Optimisation, Production automation.	Improved, optimised and new supply chain and logistics solutions. New or improved robotic systems for manufacturing and automation. New product development via manufacturing innovations.
Cluster H – Energy Management and Renewable Sources	Energy, Renewable Energy, Electrics & Electronics.	Simulation, Electricity production assessment, Wind power, Solar power, Smart grids, Autodesk, AutoCAD, PVsyst, PV/Sol, Altium Designer.	Limited development of relevant materials and equipment; primarily plan and assembly of renewable energy systems. Some smart grid-related development attempts but primarily maintenance and quality control of existing energy systems.
Cluster I – Quality Control, Assessments and Auditing	Energy, Renewable Energy, Electronics, Manufacturing, Logistics, Automotive, Defence & Military.	Safety in automation, Machinery safety, Safety standards, Site management, Compliance, Regulations.	This cluster does not have any technological trajectory as it is primarily related to assurance and quality control of the R&D and manufacturing processes.

Considering the conceptual framework of HR mining, R&D job advertisement data is examined, focusing on the sectorial skillsets and capabilities, national requirements and technological trajectories (Figure 8). HR mining results show many practical findings at sectoral, national and technological levels. At the sectoral level, it is possible to see the key capabilities that are in demand, which also indicates where the sector is heading with regard to R&D and innovation activities. At the national level, using such a technique, national competitive power can be seen in a snapshot. Finally, the job descriptors reveal technological trajectories that individual companies are targeting.

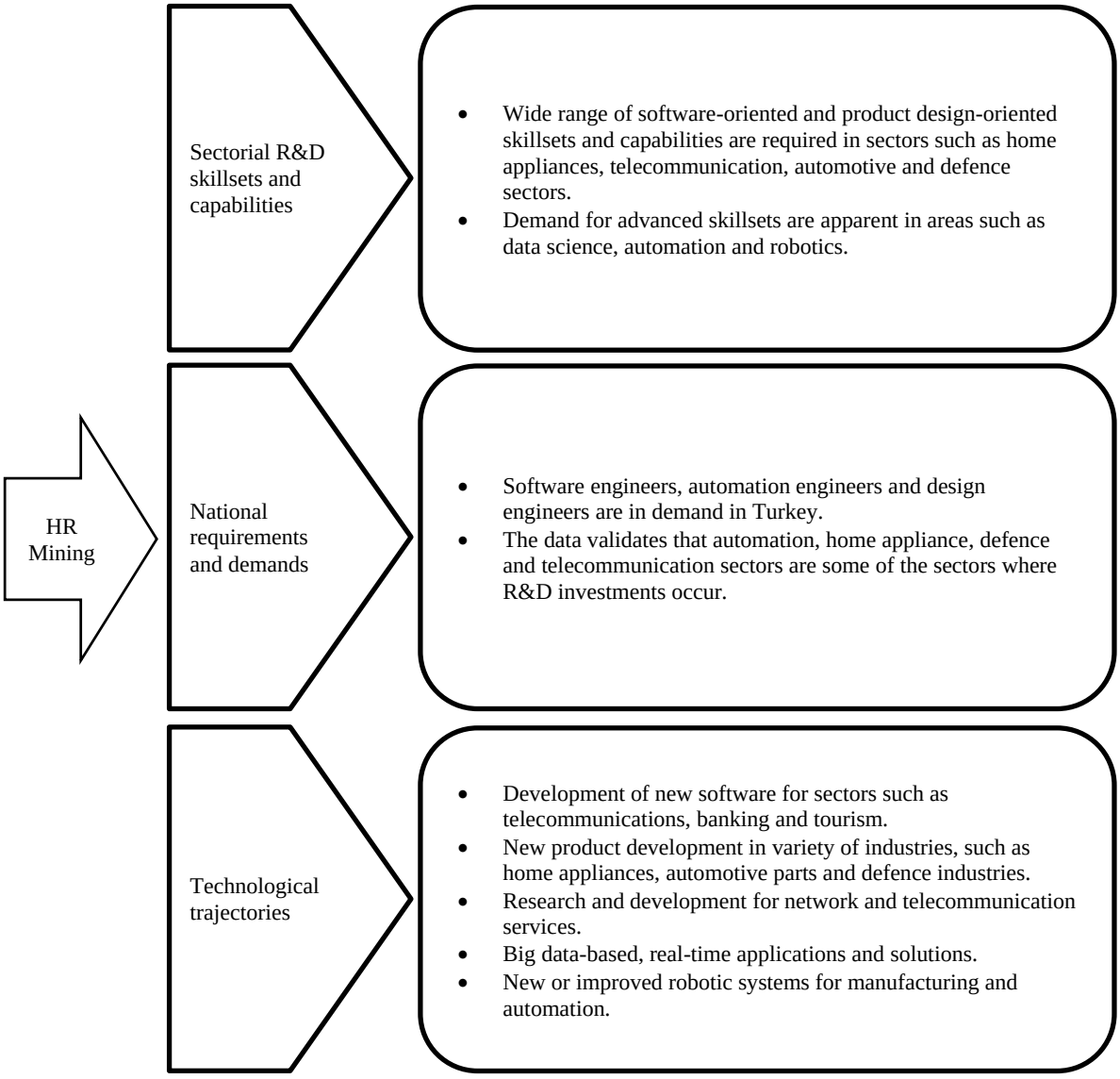


Figure 8: HR mining-based R&D skillset, sector and technology analysis

5 Conclusions

This study successfully fulfils all of its objectives. To sum, we contributed to the literature in both a methodological and practical point of view. Considering our methodological process and results, our proposed methodological approach consists of two main steps required to process job advertisements for HR mining. In the first main step of the proposed approach, we build a classification model that cleans irrelevant data from job advertisements. In this study, we focus on the R&D activities of companies, so we use the classification model to separate the R&D job advertisements from the others. The second main step is to cluster the R&D job advertisements to reveal the related keywords and skillsets in the dataset. Overall, we have applied many pre-processing operations to the text data, followed by various classification and clustering algorithms for benchmarking.

The results showed that the proposed end-to-end HR mining framework from pre-processing of the raw HR data to the clustering of the technical keywords results in interpretable results for policymakers, education sector, R&D/innovation managers and recruitment companies. Another technical contribution of our study is to show that affinity propagation algorithm, whose ability to deal with sparsity problem in short-text processing tasks has been shown before, can be successfully used to process co-occurrence matrices used to represent data such as patent documents, articles, social media contents, and job advertisements in tech mining related studies. Our results also showed that combining multiple classifiers with ensemble learning approach results in a more accurate classification model than each of the individual models.

Considering the practical contributions, our results clearly show the clusters of R&D skillsets, their inter-relations and their relevant sectors. Moreover, our analysis maps the national and the sectorial requirements and capability on an individual-skillset and R&D-projects level.

Finally, we managed to interrelate R&D job advertisements with technological and sector trajectories. We did not simply quantitatively illustrate results, but our qualitative assessments showed that some sectors are merely aiming to implement and apply what is available from other companies nationally or internationally. However, some sectors and domains (as shown in Table 2, between A-G) showed great potential to develop new products, services and businesses.

In comparison to the relevant stream of literature [49, 51-55], this study is the first to create an end-to-end HR mining approach at a national scale. The previous studies used content analysis with a smaller scale of datasets to examine job advertisement data. Our approach contributes to the literature by introducing HR mining approach with a semi-automated pipeline that can handle big data. Our study contributes to tech mining, scientometrics, bibliometrics or patentometrics literature by examining job advertisement data with the lenses of technology and R&D.

This study has practical significance and implications for policymakers, the education sector, R&D and innovation managers and recruitment companies. Policymakers and organisations in the education sector need to position themselves to fulfil what the industry requires. In contrast, investments that are outside of these results may not be highly relevant to today's conditions from a national or sectorial point of view, in Turkey. Finally, R&D and innovation management of organisations can see their sector and their competitors' investments and the sector's future direction. Accordingly, they may decide to adjust their R&D strategy and directions.

Based on the results of this study, possible applications of HR mining can be listed as shown below:

- Identification of sectoral and national skillset requirements and progress
- Assessment for technological trajectories (for the case of R&D advertisements)
- Identification of job-specific skill set requirements

- Education strategy and program development based on the HR mining results
- Competitor analysis if it is used for a particular company's job advertisement data

Our study had limitations that could be resolved with future studies. First, we focused on the methodology and our first results to align R&D job advertisements with national and sectorial trajectories. However, others may enhance this study by examining the sectorial changes over time using time series analysis. We used a job advertisement database of Turkey, and others may perform a similar examination of job advertising data at a global scale, for other nations or at a regional scale. A comparative study can be performed to show distinct national R&D capabilities indicating national competitive strengths. Finally, future studies may investigate the interrelationship between HR mining results with other examinations such as patent or publication analysis.

Acknowledgement

The authors would like to thank Kariyer.net for providing a rich database for this study. The contributions in this article by Dr. Sercan Ozcan was prepared within the framework of the Basic Research Program of the National Research University Higher School of Economics.

References

1. A. L. Porter and S. W. Cunningham, Tech mining: exploiting new technologies for competitive advantage, vol. 29. John Wiley & Sons, 2004.
2. H. Dou, V. Leveillé, S. Manullang, and D. JM Jr, "Patent analysis for competitive technical intelligence and innovative thinking," Data Science Journal, vol. 4, pp. 209–236, 2005.
3. K. M. Carvalho, E. Winter, and A. M. de Souza Antunes, "Analysis of technological developments in the treatment of alzheimer's disease through patent documents," Intelligent Information Management, vol. 7, no. 05, p. 268, 2015.
4. A. L. Porter, R. J. Watts, and T. R. Anderson, "Mining PICMET: 1997-2003 papers help you track management of technology developments," in Portland International Conference on Management of Engineering & Technology (PICMET), 2003.
5. S. K. Arora, J. Youtie, S. Carley, A. L. Porter, and P. Shapira, "Measuring the development of a common scientific lexicon in nanotechnology," Journal of Nanoparticle Research, vol. 16, p. 2194, Dec 2013.
6. X. Li, Q. Xie, L. Huang, and Z. Yuan, "Twitter data mining for the social awareness of emerging technologies," in 2017 Portland International Conference on Management of Engineering and Technology (PICMET), pp. 1–10, IEEE, July 2017.

7. N. Mikova, and A. Sokolova. "Comparing data sources for identifying technology trends." Technology Analysis & Strategic Management, vol. 31, no. 11, pp. 1353-1367, 2019.
8. A. Gok, A. Waterworth, and P. Shapira, “Use of web mining in studying innovation,” Scientometrics, vol. 102, pp.653–671, Jan 2015.
9. J. Xu, L. Guo, J. Jiang, B. Ge, and M. Li, "A deep learning methodology for automatic extraction and discovery of technical intelligence." Technological Forecasting and Social Change, vol. 146, pp. 339-351, 2019.
10. X. Shi, R. Guan, L. Wang, Z. Pei, and Y. Liang, “An incremental affinity propagation algorithm and its applications for text clustering,” in 2009 International Joint Conference on Neural Networks, pp. 2914–2919, IEEE, 2009.
11. R. Guan, X. Shi, M. Marchese, C. Yang, and Y. Liang, “Text clustering with seeds affinity propagation,” IEEE Transactions on Knowledge and Data Engineering, vol. 23, no. 4, pp. 627–637, 2011.
12. Huang, Y., Zhang, Y., Youtie, J., Porter, A.L. and Wang, X., 2016. How does national scientific funding support emerging interdisciplinary research: A comparison study of big data research in the US and China. PloS one, vol. 11, no. 5, p. e0154509, 2016.
13. K. S. Momaya and L. Lalwani, “Systems of technological innovation: a review of research activities taking the case of nanotechnology and India,” Technology Analysis & Strategic Management, vol 29, no. 6, pp. 626–641, 2017.
14. D. J. Schoeneck, A. L. Porter, R. N. Kostoff, and E. M. Berger, “Assessment of Brazil's research literature,” Technology Analysis & Strategic Management, vol 23, no. 6, pp. 601–621, 2011.
15. N. Islam and S. Ozcan, “Nanotechnology innovation system: An empirical analysis of the emerging actors and collaborative networks,” IEEE Transactions on Engineering Management, vol. 60, no. 4, pp. 687–703, 2013.
16. C. Yang, C. Huang, and J. Su, “An improved SAO network-based method for technology trend analysis: A case study of graphene,” Journal of Informetrics, vol. 12, no. 1, pp. 271–286, 2018.
17. Y. Zhang, A. L. Porter, Z. Hu, Y. Guo, and N. C. Newman, “‘Term clumping’ for technical intelligence: A case study on dye-sensitized solar cells,” Technological Forecasting and Social Change, vol. 85, pp. 26–39, 2014.
18. A. L. Porter, Y. Guo, and D. Chiavatta, 2011. “Tech mining: Text mining and visualization tools, as applied to nanoenhanced solar cells,” Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol 1, no. 2, pp. 172–181, 2011.
19. X. Zhou, A. L. Porter, D. K. Robinson, M. S. Shim, and Y. Guo, 2014. Nano-enabled drug delivery: A research profile,” Nanomedicine: Nanotechnology, Biology and Medicine, vol 10, no. 5, pp. e889–e896, 2014.
20. V. Kayser and K. Blind, “Extending the knowledge base of foresight: The contribution of text mining,” Technological Forecasting and Social Change, vol. 116, pp. 208–215, 2017.
21. F. Madani and C. Weber, “The evolution of patent mining: Applying bibliometrics analysis and keyword network analysis,” World Patent Information, vol. 46, pp. 32–48, 2016.
22. Y. Huang, J. Youtie, A. L. Porter, D. K. Robinson, S. W. Cunningham, and D. Zhu, “Big data and business: Tech mining to capture business interests and activities around big data,” in 2016 IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom) (BDCloud-SocialCom-SustainCom), pp. 145–150, IEEE, 2016.
23. A. L. Porter, “Tech mining for future-oriented technology analyses,” Futures Research Methodology, 2009.
24. V. Kayser, K. Goluchowicz, and A. Bierwisch, “Text mining for technology roadmapping—The strategic value of information,” International Journal of Innovation Management, vol. 18, no. 3, p. 1440004, 2014.
25. K. Haegeman, J. C. Harper, and R. Johnston, “Introduction to a special section: Impacts and implications of future-oriented technology analysis for policy and decision-making,” Science and Public Policy, vol. 37, no. 1, pp. 3–6, 2010.
26. F. Madani, “‘Technology Mining’ bibliometrics analysis: applying network analysis and cluster analysis,” Scientometrics, vol. 105, no. 1, pp. 323–335, 2015.

27. P. Érdi, K. Makovi, Z. Somogyvári, K. Strandburg, J. Tobochnik, P. Volf, and L. Zalányi, “Prediction of emerging technologies based on analysis of the US patent citation network,” Scientometrics, vol. 95, pp. 225–242, Apr 2013.
28. S.-B. Chang, K.-K. Lai, and S.-M. Chang, “Exploring technology diffusion and classification of business methods: Using the patent citation network,” Technological Forecasting and Social Change, vol. 76, no. 1, pp. 107–117, 2009.
29. [101] L. Zhu, D. Zhu, X. Wang, S. W. Cunningham, and Z. Wang, "An integrated solution for detecting rising technology stars in co-inventor networks." Scientometrics, vol. 121, no. 1, pp. 137-172, 2019.
30. J. , H. Park, and K. Kim, “Identifying technological competition trends for R&D planning using dynamic patent maps: SAO-based content analysis,” Scientometrics, vol. 94, pp. 313–331, 2013.
31. B. Yoon, R. Phaal, and D. Probert, “Structuring technological information for technology roadmapping: data mining approach,” in Proceedings of the 7th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases (AIKED’08), vol. 8, pp. 417–422, 2008.
32. M. N. Kyebambe, G. Cheng, Y. Huang, C. He, and Z. Zhang, “Forecasting emerging technologies: A supervised learning approach through patent analysis,” Technological Forecasting and Social Change, vol. 125, pp. 236–244, 2017.
33. C. Litecky, A. Aken, A. Ahmad, and H. Nelson, “Mining for computing jobs,” IEEE Software, vol. 27, no. 1, pp. 78–85, 2010.
34. F.-M. Tseng, C.-H. Hsieh, Y.-N. Peng, and Y.-W. Chu, “Using patent data to analyze trends and the technological strategies of the amorphous silicon thin-film solar cell industry,” Technological Forecasting and Social Change, vol. 78, no. 2, pp. 332–345, 2011.
35. B. J. Frey and D. Dueck, “Clustering by passing messages between data points,” Science, vol. 315, pp. 972–976, 2007.
36. A. Kazantseva and S. Szpakowicz, “Linear text segmentation using affinity propagation,” in Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 284–293, Association for Computational Linguistics, 2011.
37. I. Qasim, J.-W. Jeong, J.-U. Heu, and D.-H. Lee, “Concept map construction from text documents using affinity propagation,” Journal of Information Science, vol. 39, no. 6, pp. 719–736, 2013.
38. A. Rangrej, S. Kulkarni, and A. V. Tendulkar, “Comparative study of clustering techniques for short text documents,” in Proceedings of the 20th International Conference Companion on World Wide Web, pp. 111–112, ACM, 2011.
39. J. H. Kang, K. Lerman, and A. Plangprasopchok, “Analyzing microblogs with affinity propagation,” in Proceedings of the First Workshop on Social Media Analytics, pp. 67–70, ACM, 2010.
40. S. Bass and L. Kurgan, “Discovery of factors influencing patent value based on machine learning in patents in the field of nanotechnology,” Scientometrics, vol. 82, no. 2, pp. 217–241, 2009.
41. S. Venugopalan and V. Rai, “Topic based classification and pattern identification in patents,” Technological Forecasting and Social Change, vol. 94, pp. 236–250, 2015.
42. S. Li, J. Hu, Y. Cui, & J. Hu, “DeepPatent: patent classification with convolutional neural networks and word embedding. Scientometrics, vol. 117, no. 2, pp. 721–744, 2018.
43. [100] J. Ma, N. F. Abrams, A. L. Porter, D. Zhu, and D. Farrell, "Identifying translational indicators and technology opportunities for nanomedical research using tech mining: The case of gold nanostructures." Technological Forecasting and Social Change, vol. 146, pp. 767-775, 2019.
44. The U.S. Bureau of Labor Statistics, “Job openings and labor turnover summary,” https://www.bls.gov/news.release/jolts.nr0.htm, 2019.
45. The Office for National Statistics, “UK labour market: December 2018,” https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/bulletins/uklabourmarket/december2018#vacancies, 2018.
46. C. Hauff and G. Gousios, “Matching github developer profiles to job advertisements,” in Proceedings of the 12th Working Conference on Mining Software Repositories, pp. 362–366, IEEE Press, 2015.
47. G. Bal, A. Karakaş, T. Güngör, F. Süzen, and K. C. Kara, “A matching approach based on term clusters for e-recruitment,” in Industrial Conference on Data Mining, pp. 394–404, Springer, 2016.

48. I. Paparrizos, B. B. Cambazoglu, and A. Gionis, “Machine learned job recommendation,” in Proceedings of the fifth ACM Conference on Recommender Systems, pp. 325–328, ACM, 2011.
49. P. A. Todd, J. D. McKeen, and R. B. Gallupe, “The evolution of IS job skills: A content analysis of IS job advertisements from 1970 to 1990,” MIS Quarterly, vol. 19, pp. 1–27, 1995.
50. M. A. Kennan, P. Willard, D. Cecez-Kecmanovic, and C. S. Wilson, “IS knowledge and skills sought by employers: A content analysis of Australian IS early career online job advertisements,” Australasian Journal of Information Systems, vol. 15, no. 2, 2008.
51. G. L. Heimer, “Defining electronic librarianship: A content analysis of job advertisements,” Public Services Quarterly, vol. 1, no. 1, pp. 27–43, 2002.
52. L. A. Clyde, “An instructional role for librarians: an overview and content analysis of job advertisements,” Australian Academic & Research Libraries, vol. 33, no. 3, pp. 150–167, 2002.
53. R. Bennett, “Employers' demands for personal transferable skills in graduates: A content analysis of 1000 job advertisements and an associated empirical study,” Journal of Vocational Education and Training, vol. 54, no. 4, pp. 457–476, 2002.
54. R. Harper, “The collection and analysis of job advertisements: A review of research methodology,” Library and Information Research, vol. 36, no. 112, pp. 29–54, 2012.
55. S. Sanchez-Cuadrado, J. Morato, Y. Andreadakis, and J. A. Moreiro, “A study of labour market information needs through employers’ seeking behaviour,” Information Research, vol. 15, no. 4, paper 441, 2010.
56. L. Marion, M. A. Kennan, P. Willard, and C. S. Wilson, “A tale of two markets: employer expectations of information professionals in Australia and the United States of America,” in World Library and Information Congress: 71st IFLA General Conference and Council Libraries: A voyage of discovery, 2005.
57. I. Bakan, İ. F. Doğan, Competitiveness of the industries based on the Porter’s diamond model: An empirical study. International Journal of Research and Reviews in Applied Sciences, vol. 11(3), pp. 441-455, 2012.
58. P. J. Curran, Competition in UKHigher Education: Competitive Advantage in the Research Assessment Exercise and Porter's Diamond Model. Higher Education Quarterly, vol. 54(4), pp. 386-410, 2000.
59. O. Granstrand, P. Patel, K. Pavitt, Multi-technology corporations: why they have “distributed” rather than “distinctive core” competencies. California management review, vol. 39(4), pp. 8-25, 1997.
60. C. K. Prahalad, The role of core competencies in the corporation. Research-Technology Management, vol. 36(6), pp. 40-47, 1993.
61. A. Clardy, The strategic role of human resource development in managing core competencies. Human Resource Development International, vol. 11(2), pp. 183-197, 2008.
62. D. J. Teece, G. Pisano, A. Shuen, Dynamic capabilities and strategic management. Strategic management journal, vol. 18(7), pp. 509-533, 1997.
63. A. Chatterji, A. Patro, Dynamic capabilities and managing human capital. Academy of Management Perspectives, vol. 28(4), pp. 395-408, 2014.
64. M.A. Peteraf, The cornerstones of competitive advantage: a resource-based view. Strategic management journal, vol. 14(3), pp. 179-191, 1993.
65. Michael E. Porter (1990), The Competitive Advantage of Nations, New York: Free Press
66. C. Lee, O. Kwon, M. Kim, and D. Kwon, “Early identification of emerging technologies: A machine learning approach using multiple patent indicators,” Technological Forecasting and Social Change, vol. 127, pp. 291–303, 2018.
67. R. Haupt, M. Kloyer, and M. Lange, “Patent indicators for the technology life cycle development,” Research Policy, vol. 36, no. 3, pp. 387–398, 2007.
68. M. F. Porter, “Snowball: A language for stemming algorithms,” https://snowballstem.org, 2001.
69. M. Steinbach, L. Ertöz, and V. Kumar, “The challenges of clustering high dimensional data,” in New Directions in Statistical Physics, pp. 273–309, Springer, 2004.
70. L. Liu, J. Kang, J. Yu, and Z. Wang, “A comparative study on unsupervised feature selection methods for text clustering,” in 2005 International Conference on Natural Language Processing and Knowledge Engineering, pp. 597–601, IEEE, 2005.

71. A. Huang, “Similarity measures for text document clustering,” in Proceedings of the Sixth New Zealand Computer Science Research Student Conference (NZCSRSC 2008), pp. 49–56, 2008.
72. S. Ranaei, A. Suominen, A. Porter, and T. Kässi. "Application of Text-Analytics in Quantitative Study of Science and Technology." In Springer Handbook of Science and Technology Indicators, pp. 957-982. Springer, Cham, 2019.
73. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, “Scikit-learn: Machine learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
74. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: synthetic minority over-sampling technique,” Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.
75. S. Maldonado, J. López, and C. Vairetti, “An alternative SMOTE oversampling strategy for high-dimensional datasets,” Applied Soft Computing, vol. 76, pp. 380–389, 2019.
76. A. Genkin, D. D. Lewis, and D. Madigan, “Large-scale bayesian logistic regression for text categorization,” Technometrics, vol. 49, no. 3, pp. 291–304, 2007.
77. W. S. Lee and B. Liu, “Learning with positive and unlabeled examples using weighted logistic regression,” in ICML, vol. 3, pp. 448–455, 2003.
78. C. Cortes and V. Vapnik, “Support-vector networks,” Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.
79. E. Bauer and R. Kohavi, “An empirical comparison of voting classification algorithms: Bagging, boosting, and variants,” Machine Learning, vol. 36, no. 1-2, pp. 105–139, 1999.
80. J. H. Friedman, “Greedy function approximation: a gradient boosting machine,” Annals of Statistics, pp. 1189–1232, 2001.
81. T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD’16, pp. 785–794, 2016.
82. O. Sagi and L. Rokach, “Ensemble learning: A survey,” Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 8, no. 4, p. e1249, 2018.
83. T. G. Dietterich, “Ensemble methods in machine learning,” International Workshop on Multiple Classifier Systems, pp. 1–15, Springer, 2000.
84. A. Saxena, M. Prasad, A. Gupta, N. Bharill, O. P. Patel, A. Tiwari, M. J. Er, W. Ding, and C.-T. Lin, “A review of clustering techniques and developments,” Neurocomputing, vol. 267, pp. 664–681, 2017.
85. Y. P. Raykov, A. Boukouvalas, F. Baig, and M. A. Little, “What to do when k-means clustering fails: A simple yet principled alternative algorithm,” PLOS One, vol. 11, no. 9, p. e0162259, 2016.
86. I. S. Dhillon and D. S. Modha, “Concept decompositions for large sparse text data using clustering,” Machine Learning, vol. 42, pp. 143–175, Jan. 2001.
87. D. Arthur and S. Vassilvitskii, “k-means++: The advantages of careful seeding,” in Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 1027–1035, 2007.
88. G. Carlsson and F. Mémoli, “Characterization, stability and convergence of hierarchical clustering methods,” Journal of Machine Learning Research, vol. 11, pp. 1425–1470, 2010.
89. M. S. G. Karypis, V. Kumar, and M. Steinbach, “A comparison of document clustering techniques,” in KDD Workshop on Text Mining, 2000.
90. M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, “A density-based algorithm for discovering clusters in large spatial databases with noise,” in KDD, 1996.
91. G. Gan, C. Ma, and J. Wu, Data clustering: theory, algorithms, and applications, SIAM, 2007.
92. H. Jiang, J. Li, S. Yi, X. Wang, and X. Hu, “A new hybrid method based on partitioning-based DBSCAN and ant clustering,” Expert Systems with Applications, vol. 38, no. 8, pp. 9373–9381, 2011.
93. U. Maulik and S. Bandyopadhyay, “Performance evaluation of some clustering algorithms and validity indices,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 12, pp. 1650–1654, 2002.
94. P. J. Rousseeuw, “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis,” Journal of Computational and Applied Mathematics, vol. 20, pp. 53–65, 1987.

95. J. C. Dunn, “Well-separated clusters and optimal fuzzy partitions,” Journal of Cybernetics, vol. 4, no. 1, pp. 95–104, 1974.
96. Y. Jia, J. Hoberock, M. Garland, and J. Hart, “On the visualization of social and other scale-free networks,” IEEE Transactions on Visualization and Computer Graphics, vol. 14, no. 6, pp. 1285–1292, 2008.
97. D. Auber, Y. Chiricota, F. Jourdan, and G. Melanc¸on, “Multiscale visualization of small world networks,” IEEE Symposium on Information Visualization 2003, pp. 75–81, 2003.
98. M. Jacomy, T. Venturini, S. Heymann, and M. Bastian, “Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software,” PLOS One, vol. 9, no. 6, p. e98679, 2014.
99. J. Lewis, M. Ackerman, and V. de Sa, “Human cluster evaluation and formal quality measures: A comparative study,” in Proceedings of the 34th Annual Conference of the Cognitive Science Society (CogSci), pp. 1870–1875, 2012.